BEST AVAILABLE COPY

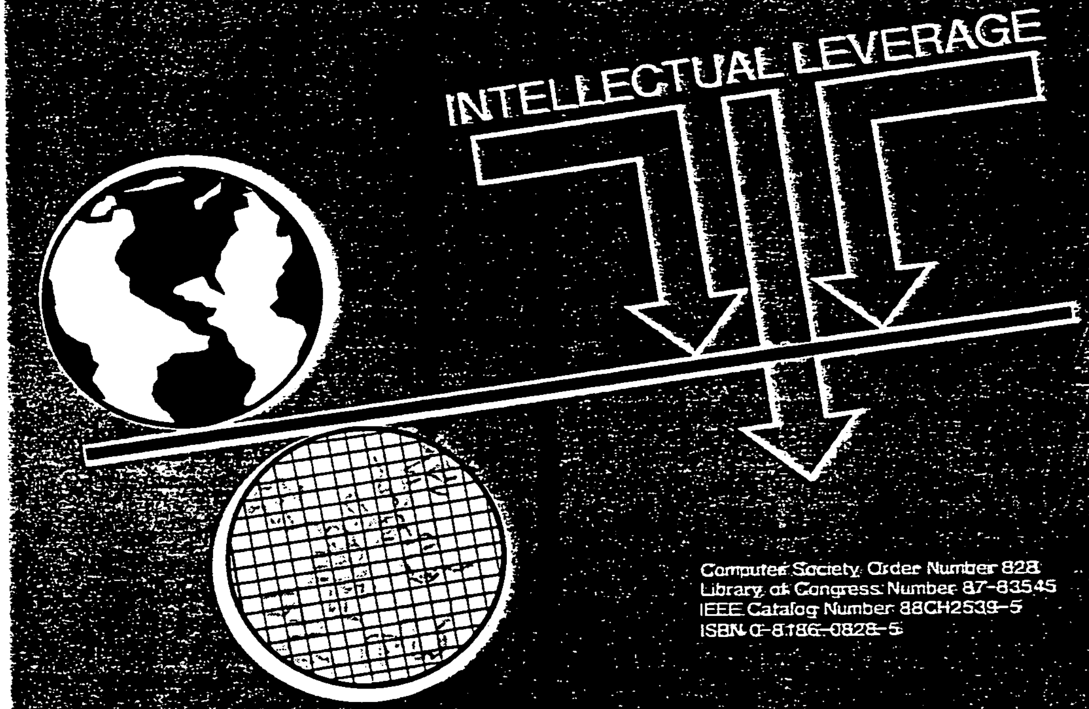
digest of papers

February 29 – March 4, 1988

spring COMPCON 88

THIRTY-THIRD IEEE COMPUTER SOCIETY INTERNATIONAL CONFERENCE
CATHEDRAL HILL HOTEL, SAN FRANCISCO, CALIFORNIA

San Francisco



Computer Society Order Number 828
Library of Congress Number 87-83545
IEEE Catalog Number 88CH2539-5
ISBN 0-8186-0828-5

THE COMPUTER SOCIETY OF THE IEEE
THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.
COMPUTER SOCIETY PRESS

The IBM 3990 Disk Cache

Jai Menon

IBM Almaden Research Center
San Jose, California 95120-6099

Mike Hartung

IBM Tucson Laboratory
Tucson, Arizona 85744

Abstract

This report gives a brief overview of the IBM 3990 Model 3 cache control unit. Some new performance and RAS features are described and some modelling results on performance are presented.

Introduction

Disk caches that are used to extend the performance of high-end computer systems have been implemented by several vendors (see [SMITH85] for a list of some of these vendors). This paper describes the IBM 3990 cached disk control unit, which is the most modern IBM disk cache to date ([IBM3990]).

The 3990 is available in three models ([IBM3990]), and the structure and features of the 3990 family of control units represent a significant improvement over the structure and features of the older 3880 family of control units ([IBM3880, IBM232]). The 3990 Model 1 and the 3990 Model 2 are uncached control units, and we will not describe them any more in this report. The 3990 Model 3 is a cached disk control unit, and the rest of this paper will be devoted to a description of the 3990 Model 3. For simplicity, in the rest of this report, we will often use the term "3990" when we really mean the "3990 Model 3". We will also sometimes refer to a disk as DASD (direct access storage device), which is the IBM terminology for disk.

The 3990 is a separate, stand-alone control unit, which attaches to the host CPU over ***channels*** and to the IBM 3380 disk via an internally defined disk interface. For an overview of a typical I/O in IBM architecture, and a summary of what functions are typically performed in a channel and what functions are typically performed in a control unit, the reader is referred to [BOHL81].

Frequently referenced records from the disk are stored in high-speed electronic storage in the 3990. When the host CPU asks for a record from a disk, the 3990 first checks to see if the requested record is in the electronic storage (cache) in the 3990. If so, the record is returned from the cache and there is no need to access the disk. The more often requested data is found in the cache (a ***cache hit*** occurs), the better the performance of the control unit. If a copy of the record is in the cache when the host initiates a read request, we will refer to that as a ***read hit***. On the other hand, if a copy of the record is in the cache when the host initiates a write request, we will call that a ***write hit***.

The rest of this paper will be organized as follows. We will begin with a diagram and overview of the paths in the 3990. Then, we will describe the performance and cache features of the 3990. Following this, we describe the dual copy feature and other reliability, availability and serviceability features of the 3990. Finally, we will present some results indicating the expected performance that can be obtained by using a 3990 disk cache.

3990 Overview

In Figure 1 on page 2, we show an overview of the 3990 Model 3. As can be seen, it consists of two independent ***storage clusters***. Each cluster provides a separate power and service region and two separate paths to the string of DASDs. Loss of power to one cluster does not disable the 3990, since processing continues through the other storage cluster.

CH2539-5/88/0000/0146$01.00 © 1988 IEEE

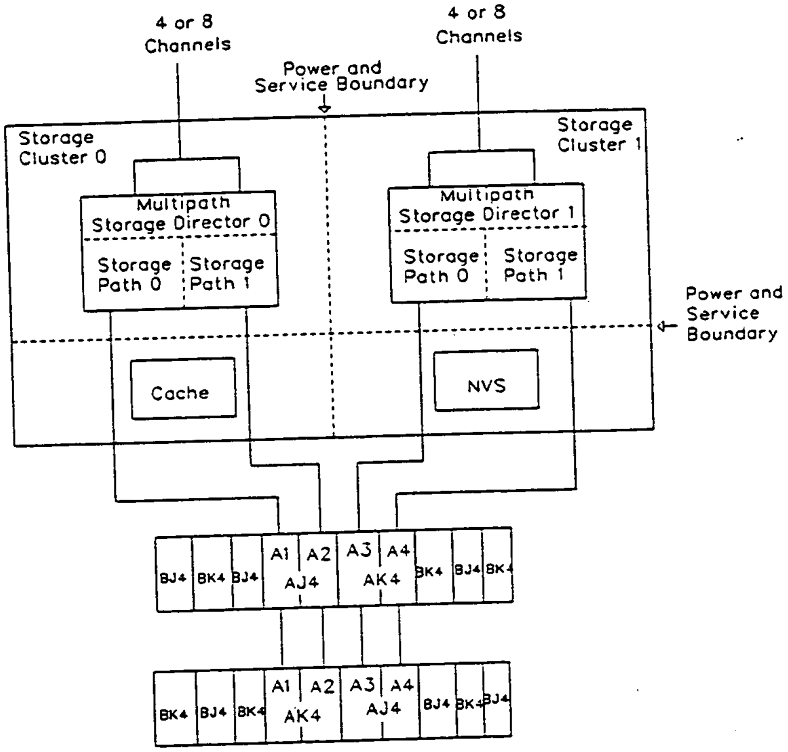


Figure 1: The IBM 3990 Disk Cache

Each storage cluster attaches to host CPUs via four or eight channels. Each of the possible total of 16 channels can operate at either 3 or 4.5 Mbytes/sec.

As a new standard of data availability and overall performance, twice as many storage paths are provided in the 3990 than in the older 3880 Model 23 ([IBM231, IBM232]). When attached to the new 4-path 3380 Models AJ4/AK4 ([IBM3380]), the 3990 can access any device over any one of ***four*** paths using the ***device level selection enhanced*** (DLSE) mode. To repeat, DLSE provides four paths to each actuator and simultaneous data transfer to any four actuators in the attached 4-path strings of devices. In Figure 1, each AJ4, AK4, BJ4 or BK4 actually represents four devices (actuators). Thus, a total of 64 devices are represented in the figure. The J and K in AJ4, AK4, BJ4 and BK4 refer to different capacity 3380 disks, and, as is apparent from Figure 1, any combination of the J and K disks may be intermixed. For further details on these devices, and the difference between the AJ4 (AK4) and the BJ4 (BK4), the reader is referred to [IBM3380].

The 3990 has cache sizes of 32, 64, 128 or 256 Mbytes. Cache is in a separate power region from the storage clusters. If a storage cluster is off-line, cache processing still continues through the other storage cluster.

Finally, the 3990 has nonvolatile storage (NVS) which provides random-access electronic storage. However, as will become apparent later in the paper, the NVS is not used as a cache. The NVS has its own separate power region for data protection and is used to perform functions like DASD fast write and dual copy, which will be described later. If power is lost to the 3990 before certain data in NVS has been copied to disk, a battery-backup system maintains power in the nonvolatile storage for up to 48 hours with a fully-charged battery to prevent data loss. When power is restored, the 3990 destages any data in NVS to disk and completes any operation in process at the time of the failure.

Details of the 3990 Cache

The cache in the 3990 which is shared by all the storage paths uses IBM's one million bit DRAM chip. At any time, there may be as many as four different devices that are sending data to be placed in the cache, and there may be as many as four different channels that are accessing data from the cache. Thus, the cache permits eight simultaneous accesses. This capability of the 3990 cache to allow four simultaneous and independent operations to/from the channels and, at the same time, four independent and simultaneous transfers to/from the devices is called the ***dual data transfer*** capability. A least-recently-used algorithm (and other algorithms) keeps high-activity data in the cache because it has the highest probability of reuse.

The 3990 cache is divided into 16K byte segments. When a specified amount of cache space is needed, as many segments as are needed to hold the specified object are allocated. These allocated segments need not be contiguous. The 3990 can logically relate separated segments in the cache and treat them as a single unit of data.

Data transfers between the cache and the channel operate at the maximum speed of the channel, either 3 or 4.5 Mbytes/sec. All data transfers requiring access to the disk will occur at the disk transfer rate of 3 Mbytes/sec.

Cache Operations

With the 3990, there are three different modes for cache operations:

- Basic caching
- DASD fast write
- Cache fast write

With basic caching, only read operations benefit from cache. This mode of caching was the only one provided in the earlier IBM 3880 control unit. The 3990 provides two new modes of caching operations - ***DASD fast write*** and ***cache fast write***. For simplicity, we will refer to a write operation executed in DASD fast write mode as a DASD fast write and to a write operation executed in cache fast write mode as a cache fast write. With DASD fast write and cache fast write, the performance benefits of caching are extended to write operations.

The I/O command must specify which of the three modes of caching is to be employed. Read operations are executed similarly in all three caching modes and will be described first. Write operations are executed differently depending on the mode, as will become clear from the following discussion.

Execution of Reads in All Three Modes

If a copy of the data is in the cache when the host initiates a read request (read hit), the 3990 transfers the desired data from the cache to the channel. If a copy of the data is not in cache (read miss), the 3990 sends the requested data directly to the channel from the disk and, at the same time, writes that data (plus the rest of the data from that record to the end of the track) into the cache in anticipation of future use. This type of operation is called a ***branching transfer***, since data being read from the disk is simultaneously branched to two destinations (the channel and the cache). Future requests for the referenced record or for following records on that track are read from the cache and are read hits.

Execution of Writes in Basic Caching Mode

If a copy of the data is in the cache when the host initiates a write operation in basic caching mode, the 3990 writes the data directly to DASD and, at the same time, writes that data into the cache (a branching transfer is used for this purpose also). The record in cache is updated because it may be referred to again. However, before operation complete can be returned, the 3990 also ensures that the record has been successfully written to DASD. On a write miss, the record is directly written to DASD and not written to cache.

It should be clear from this discussion that whether a copy of the record is in the cache or not, the total I/O time for the write operation in basic caching mode is approximately the same as for uncached control units, and is governed by the actual disk access and transfer times.

Execution of Writes in the Two Fast Write Modes

Next, let us describe the operation of the 3990 cache in executing write operations in the two fast write modes. Both types of fast write operations - DASD fast write and cache fast write can improve performance for write hits or full track format write operations. Most write operations are hits because typical applications read a record before updating it, or the write operation itself creates a new record. This last operation is called a ***format write***. In a format write, the new record is written, and the rest of the track is formatted for new data. Thus, there is no need to verify the data on the track before allowing the cache write. Hence, format writes can be considered as cache hits.

Unlike write operations in the basic caching mode, fast writes use NVS as will be described shortly.

DASD Fast Write

DASD fast write improves storage subsystem performance because immediate access to DASD is not required for

WORK COPY

write hits and full track format writes. DASD fast write stores data simultaneously in cache and in nonvolatile storage using a branching transfer. Access to DASD is not required to complete the DASD fast write operation. Because a copy of the data is put into the NVS, the 3990 returns operation complete at the end of data transfer to cache and NVS. This allows the program in the host CPU to continue processing without waiting for the data to be put on DASD. The data remains in cache and in NVS until the data is written to DASD to free space in the cache or NVS. Thus, most write operations operate directly with the cache and NVS without going to DASD, resulting in the same performance as a read hit.

On a write miss in DASD fast write mode, the 3990 writes the data to DASD and cache simultaneously. The remainder of the track is staged into the cache.

So completion [illegible]
[illegible] written [illegible]
[illegible] [illegible] cache

Cache Fast Write

Cache fast write is an option of the 3990 designed for use with special kinds of data, such as temporary work files produced during sorting. Such data does not need to be written to disk, and the I/O is considered complete as soon as data is written into the cache. Unlike DASD fast write, a copy of the data is not written into NVS. The data may never get written to disk.

On a write miss in cache fast write mode, the 3990 writes the data to DASD and cache simultaneously. The remainder of the track is staged into the cache. Thus, write misses work similarly in DASD fast write and cache fast write modes.

Cache Algorithms

Caching algorithms include ***normal, sequential, bypass-cache*** and ***inhibit cache loading***. Normal caching algorithms are used unless otherwise directed in the I/O command. This means that data is staged to the cache after being referred to in a read operation and remains in the cache until LRU algorithms permit the data to be overlaid by other data. During sequential caching, the 3990 pre-stages anticipated data so that up to five tracks are in the cache. Bypass-cache does not use the cache and operations go directly to the disk. Inhibit cache loading uses existing copies of tracks if they are in the cache, but does not load any new tracks into the cache.

Dual Copy in the 3990

Dual copy allows the 3990 to maintain a duplicate copy of the data on a device on a different device. This improves the availability of data. The status of the dual copy operation is kept in NVS.

The two physical devices are a ***duplex pair*** - a primary device and a secondary device. The dual copy operation is managed by the 3990. All I/O operations are directed to the primary device. The 3990 automatically updates both copies of the data. Data is accessed from the secondary device if the primary device is not available.

Because the secondary device is off-line, the host knows only of one device - the primary device. The 3990 orients to the primary device and does a branching transfer of the data to the primary device and to the cache. It also updates status information in the NVS to indicate that the primary device has been updated and that the secondary device has not yet been updated. At this point, the 3990 returns operation complete to the host. Later, the 3990 completes the write operation from the cache to the secondary device. This write operation is transparent to the host.

The DASD fast write capability and the dual copy capability can be combined to form a ***fast dual copy***. Using fast dual copy results in an optimum of data availability, performance and reliability. The method of operation for fast dual copy depends on whether a write hit or a write miss occurs. On a write miss, the operation is similar to that for dual copy described above. For fast dual copy with write hit, the 3990 does a branching transfer of the data to cache and NVS. It then updates status information in the NVS regarding the state of the dual copy operation. At this point, the 3990 returns operation complete to the host. Later, the 3990 completes the write operation from the cache to the primary device, and, later still, from the cache to the secondary device.

RAS Features on the 3990

The 3990 provides a number of reliability, availability and serviceability (RAS) features. A major improvement in RAS for the storage subsystem is provided by dual copy, which we have already described. A second RAS feature is the use of a pair of independent storage clusters in the 3990. A clear benefit of each of the clusters is that they are independent components with separate power and service regions. Each storage cluster has its own support facility. A major RAS enhancement, the support facility permits concurrent maintenance and provides a remote maintenance support capability. Among other tasks, the support facility monitors subsystem activity, generates service information messages (SIMs), communicates with the other support facility, runs maintenance analysis programs and diagnostics, and logs error conditions on diskette storage.

As a consequence of concurrent maintenance, one storage cluster can continue to access cache and DASD while maintenance activities are taking place on the other storage cluster. Also, a service action can be performed on the cache while direct access to DASD is provided through the storage clusters, and a service action can be performed on NVS while caching operations and direct access to DASD continues through the storage clusters.

The remote maintenance support capability permits a support representative to establish communication with either storage cluster through an external modem. Once established, the remote service representative can analyze the error data and send maintenance information to the service representative on site.

The 3990 has a writeable diskette which contains 3990 microcode, microcode patches, error log, and maintenance analysis procedures. During either local or remote maintenance, microcode patches can be transmitted to the 3990 support facility and stored on the 3990's diskette. Microcode patches written on the diskette are not lost across IMLs and are not installed until the installation asks a local service representative to do so.

Performance Modelling Results for the 3990

In this section, we present performance results from mathematical models. The different operating environments and processing workloads used with the model were obtained through studies of representative production systems. Some of these models have not yet been validated against 3990s running in a production environment, so the results must be taken with appropriate caution.

For the studies, we modelled a single 3990 with 32 Mbytes of cache, and 32 devices operating in the TSO environment. For this environment, the average block size has been measured to be 5496 bytes, the read hit ratio to be 91% and the write hit ratio to be 98%.

Performance numbers are typically given in terms of I/Os per second at a given response time. At 22 msecs response time, we found that the given configuration could provide 500 I/Os per second. Assuming that fast dual copy was performed on all 32 devices (in other words, there were physically 64 devices), the I/O rate dropped to 380 per second at the same 22 msec response time. Finally, with DASD fast write, the given configuration of 32 devices could provide 590 I/Os per second. These results were obtained assuming 4.5 Mb/sec channels. When these numbers were compared to a similar configuration using the older 3880 Model 23 disk cache, we found that the 3990 provided 20% higher throughput using fast dual copy and 90% higher throughput using DASD fast write.

Next, we compared the response times at a fixed I/O rate of 150 per second. The 3880 Model 23 provided 8.5 msecs response time, the 3990 provided 6.0 msecs response time, which dropped to 4.2 msecs with fast dual copy and 4.0 msecs with DASD fast write.

Conclusions

The IBM 3990 disk cache is designed to meet and exceed the data availability, performance and reliability requirements that businesses demand today. To improve performance, it uses cache sizes as large as 256 Mbytes, uses improved cache slot segmentation resulting in more efficient cache space utilization, provides two new fast write capabilities - DASD fast write and cache fast write, allows for branching and dual data transfers, and provides four paths to DASD. To improve RAS, it provides a dual copy capability, two independent storage clusters, separate power and service regions for cache and nonvolatile storage and concurrent and remote maintenance support. Overall, the 3990 represents a new dimension in disk caching.

Bibliography

[BOHL81] Bohl, M., Introduction to IBM Direct Access Storage Devices, SRA (1981).

[IBM231] IBM 3880 Storage Control Model 23 Description, *IBM Publication* GA32-0083 (Tucson, 1985).

[IBM232] IBM 3880 Storage Control Model 23 Introduction, *IBM Publication* GA32-0082 (Tucson, 1985).

[IBM3380] IBM 3380 Direct Access Storage Introduction, *IBM Publication* GC26-4491-0.

on of 32 devices
nese results were
When these num-
uration using the
nd that the 3990
ıst dual copy and
ıt write.

t a fixed I/O rate
rovided 8.5 msecs
ecs response time,
ual copy and 4.0

meet and exceed
eliability require-
ı improve perfor-
!56 Mbytes, uses
g in more efficient
r fast write capa-
. write, allows for
ovides four paths
des a dual copy
lusters, separate
onvolatile storage
support. Overall,
disk caching.

ibliography

IBM Direct Ac-
\ (1981).

ol Model 23 De-
tion GA32-0083

l Model 23 Intro-
ion GA32-0082

torage Introduc-
26-4491-0.

[IBM3880] IBM 3880 Storage Control Models 1, 2, 3 and 4 Description Manual, *IBM Publication* GA26-1661.

[IBM3990] IBM 3990 Storage Control Introduction, *IBM Publication* GA32-0098-0.

[SMITH85] Smith, A. J., Disk Cache - Miss Ratio Analysis and Design Considerations, *ACM Transactions on Computer Systems* 3 (Aug. 1985) pp. 161—203.